UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Text Mining for Social Networks Sentiment Analysis. Two case studies: "FIFA World Cup 2018" and "Cristiano Ronaldo signs for Juventus"

Miguel Antunes Sampaio Rodrigues

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Nuno Pombo
Coorientador: Prof. Doutor Nuno M. Garcia

Covilhã, Janeiro de 2019

Folha em branco

Dedicatória

À minha avó Celeste.

Folha em branco

Agradecimentos

No final da realização da minha dissertação, tenho os seguintes agradecimentos a fazer: em primeiro lugar à minha família, nomeadamente aos meus pais e ao meu irmão, que sempre se mostraram interessados, dispostos a ajudar e a motivar ao longo do meu percurso académico; quero agradecer também aos meus colegas de mestrado Sérgio Silva e Tiago Mendes, que igualmente se mostraram dispostos a ajudar no que fosse preciso; por fim quero agradecer ao meu orientador de mestrado, Professor Doutor Nuno Pombo, e ao meu coorientador, Professor Doutor Nuno M. Garcia que durante a realização desta dissertação sempre se mostraram disponíveis para ajudar.

Folha em branco

Resumo

Numa sociedade cada vez mais ligada pelas redes sociais é importante saber ao certo o que aí se discute, tanto sobre marcas específicas ou produtos, como sobre eventos importantes que ocorram. Esse conhecimento é de extrema importância para as entidades responsáveis pelos produtos ou eventos em questão que querem melhorar a sua prestação com o intuito de agradar cada vez mais aos utilizadores. Isto pode ser feito de uma maneira muito fácil através da monitorização de redes sociais que, como foi dito, são um meio de comunicação em crescimento constante.

A presente dissertação insere-se no contexto da análise de sentimentos baseada na extração de informação das redes sociais. Com o objetivo de se validar o modelo proposto, dois casos de estudo foram considerados: análise da competição Mundial FIFA 2018 de futebol e o "efeito Cristiano Ronaldo".

Esta dissertação pretende fazer o estudo de conteúdos da rede social *Twitter*, não se restringindo apenas ao que é escrito, mas focando-se também nas emoções que são expressas nessas publicações.

São referidas nesta dissertação, relativamente ao primeiro caso de estudo, as quatro seleções que chegaram mais longe no campeonato do mundo de futebol 2018, que foram: França, Croácia, Bélgica e Inglaterra. Além destas seleções é mencionada a de Portugal, uma vez que foi a vencedora da competição *UEFA Euro 2016*. Por fim, é feito o levantamento das emoções expressas pelos utilizadores em relação a cada seleção em particular, além do levantamento das emoções provocadas em cada jogo entre estas seleções.

Relativamente ao segundo caso de estudo, é referida a transferência do jogador Cristiano Ronaldo e ainda os clubes Real Madrid CF e Juventus FC que são o seu antigo e atual clubes, respetivamente.

Apesar de já existirem bastantes trabalhos realizados nesta área, tal não acontece concretamente sobre os dois casos de estudo propostos, o que torna mais interessante o tema desta dissertação.

No final da realização desta dissertação, relativamente ao primeiro caso de estudo, podemos concluir que a seleção que provoca sentimentos mais positivos, dentro do conjunto de todas as seleções, é a seleção francesa. Isto devido à vitória da mesma no mundial de futebol de 2018. Ainda no primeiro caso de estudo relativamente à identificação das emoções demonstradas, podemos concluir que as emoções predominantes são a de "felicidade", seguida pela emoção de "surpresa". Esta última pode ser explicada pela chegada da seleção croata à final da competição, que não fazia parte do grupo das seleções teoricamente favoritas.

Relativamente ao segundo caso de estudo, na identificação de emoções, com o objetivo de perceber como é que a transferência do jogador Cristiano Ronaldo afetou os adeptos afetos às duas equipas, concluímos que, relativamente aos adeptos da Juventus, foi predominante o

sentimento de “surpresa”, porque esta transferência não parecia ser possível. Relativamente aos adeptos do Real Madrid, a emoção predominante foi também de “surpresa”, pelas mesmas razões dos adeptos da Juventus. No entanto, foi identificado um número significante de *tweets* que expressaram “tristeza” e “medo”. Isto devido ao jogador ser considerado por muitos, o melhor jogador do mundo.
Adicionalmente nesta dissertação, é apresentado um modelo constituído por etapas, que propõe uma diretriz para a realização deste tipo de projetos, relativos à Mineração de Texto e Análise de Sentimentos.

Palavras-chave

Mineração de Texto, Mineração de Dados, *Twitter*, Processamento de Linguagem Natural, Mundial FIFA 2018, Cristiano Ronaldo, Juventus FC, Real Madrid CF

Abstract

In a society increasingly linked by social networks, it is important to know for sure what is said there, both on specific brands or products, and on important events that occur. This knowledge is extremely important for the entities responsible for the products or events in question that want to improve their performance in order to please the users more and more. This can be done very easily by monitoring social networks which, as has been said, are a constantly growing medium of communication.

The present dissertation is inserted in the context of the sentiment analysis based on the extraction of information from social networks. In order to validate the proposed model, two case studies were considered: analysis of the FIFA 2018 World Cup competition and the Cristiano Ronaldo move from Real Madrid CF to Juventus FC.

This dissertation intends to study the contents of the social network Twitter, not only being restricted to what is written, but also focusing attention on the emotions that are expressed in those publications.

In this dissertation, in relation to the first case study, the four teams that went the furthest in the competition were: France, Croatia, Belgium and England. In addition to these selections Portugal is mentioned since it was the winner of the UEFA Euro 2016 competition. Finally, it is made the survey of the emotions expressed by the users in relation to each particular selection, as well as the emotions raised in each game between these selections.

Regarding the second case study, the Cristiano Ronaldo player is mentioned, as well as the clubs Real Madrid CF and Juventus FC which are his old and current clubs, respectively.

Although there are already many researches done in this area, this does not happen concretely on the two proposed case studies, which makes the topic of this dissertation more interesting.

At the end of this dissertation, in relation to the first case study, we can conclude that the the national team that provokes more positive sentiment, within the set of all national teams, is the French national team. This is due to the victory in the FIFA World Cup 2018. Still in the first case study, regarding the emotion detection, we can conclude that the predominant emotions are the one of “happiness”, followed by the emotion of “surprise”. This last can be explained by the presence of the Croatian national team in the final, that was not part of the group of the national teams theoretically favorites.

Regarding the second case study, in the emotion detection, in order to understand how the move of the player Cristiano Ronaldo affected the fans of both teams, we concluded that, relatively, the predominant emotion was also “surprise”, for the same reason as the Juventus fans. However, a significant number of tweets expressing “sadness” and “fear” were identified. This is due to the player being considered by many, the best player in the world. Additionally, in this dissertation, is presented a model consisting of stages, which proposes a

guideline for the accomplishment of this type of projects, related to Text Mining and Sentiment Analysis.

Keywords

Text Mining, Data Mining, *Twitter*, Natural Language Processing, FIFA World Cup 2018, Cristiano Ronaldo, Juventus FC, Real Madrid CF

Folha em branco

Índice

Dedicatória iii
Agradecimentos v
Resumo vii
Abstract ix
Lista de Figuras xv
Lista de Tabelas xvi
Lista de Acrónimos xvii
1. Introdução 19
1.1 Objetivo 19
1.2 Motivação 20
1.3 Questões de Pesquisa 20
1.4 Enquadramento 20
1.5 Estrutura da dissertação 21
2. Estado da Arte 23
2.1 Análise de Sentimentos 23
2.1.1 Importância da Análise de Sentimentos 23
2.1.2 Vantagens da Análise de Sentimentos 23
2.1.3 Tipos de Análise de Sentimentos 24
2.1.4 Aplicações práticas 24
2.1.5 Métodos de Classificação de Sentimentos 25
2.1.6 Avaliação do Sentimento 25
2.2 Análise de Sentimentos no Twitter 26
2.3 Mineração de Texto 26
2.3.1 Etapas da Mineração de Texto 27
2.4 Mineração de Texto VS. Mineração de Dados 27
2.5 Revisão da Literatura 28
3. Tecnologias implementadas 31
3.1 Linguagem de programação para obter os *datasets* 31
3.1.1 Java 31
3.1.2 Python 31
3.2 *Framework* para análise de sentimentos 31
3.2.1 Orange 3 32
4. Implementação 33
4.1 Python 33
4.1.1 Extensões *Python* 33
4.2 Implementação *Python* 34
4.3 Orange 3 35
4.3.1 *Add-ons* e funcionalidades de mineração de texto 35
4.3.1.1 *Sentiment Analysis* 35
4.3.1.2 *Tweet Profiler* 35

4.3.1.3 *Distributions* 35
4.3.1.4 *Corpus* 35
4.3.1.5 *Word Cloud* 35
4.3.1.6 *Preprocess Text* 36
4.3.1.7 *Topic Modelling* 36
4.3.2 Funcionalidade de visualização e manipulação de dados 36
4.3.2.1 *Data Table* 36
4.3.2.2 *Select Columns* 36
4.3.2.3 *Data Sampler* 36
4.3.2.4 *Discretize* 37
4.3.2.5 *Scatter Plot* 37
4.3.3 Funcionalidades de classificação de dados 37
4.3.3.1 *Neural Network* 37
4.3.3.2 *Naive-Bayes* 38
4.3.3.3 *Knn* 38
4.3.3.4 *Logistic Regression* 39
4.3.3.5 *Support Vector Machine (SVM)* 40
4.3.4 Funcionalidades de avaliação 40
4.3.4.1 *Test & Score* 41
4.4 Implementação *Orange 3* 41
5. Resultados obtidos 43
5.1 Resultados da tecnologia *Orange 3* 43
5.1.2 Resultados relativos ao caso de estudo "Mundial FIFA 2018" 43
5.1.2.1 Qual a equipa cujos comentários dos adeptos têm uma polaridade mais positiva? 43
5.1.2.2 Como é que os jogos da competição *FIFA World Cup 2018* afetaram os comentários dos utilizadores e o tipo de emoções que estes expressaram? 47
5.1.3 Resultados relativos ao caso de estudo da transferência do jogador Cristiano Ronaldo do Real Madrid CF para a Juventus FC 50
5.1.3.1 Como é que a transferência do jogador Cristiano Ronaldo do Real Madrid para a Juventus afetou os comentários dos utilizadores afetos aos dois clubes? 50
6. Conclusão e trabalho futuro 55
6.1 Conclusão 55
6.2 Trabalho Futuro 56
Referências 57
Anexos 59
Anexo A – Código *Python* para o *download* dos *datasets* 59
Anexo B – Workflow do Orange 3 60
Anexo C – Word Clouds 61

Folha em branco

Lista de Figuras

Figura 2. 1 Etapas da Mineração de Texto 27
Figura 4. 1 Etapas da Implementação 33
Figura 4. 2 Rede Neuronal 38
Figura 4. 3 kNN 39
Figura 4. 4 Regressão Logística 40
Figura 4. 5 Support Vector Machine 40
Figura 5. 1 Gráfico de dispersão seleção francesa 43
Figura 5. 2 Test&Score seleção francesa 44
Figura 5. 3 Gráfico de dispersão seleção croata 44
Figura 5. 4 Test&Score seleção croata 44
Figura 5. 5 Gráfico de dispersão seleção belga 45
Figura 5. 6 Test&Score seleção belga 45
Figura 5. 7 Gráfico de dispersão seleção inglesa 45
Figura 5. 8 Test&Score seleção inglesa 46
Figura 5. 9 Gráfico de dispersão seleção portuguesa 46
Figura 5. 10 Test&Score seleção portuguesa 46
Figura 5. 11 Gráfico de barras de emoções França vs. Croácia 47
Figura 5. 12 Test&Score França vs Croácia 48
Figura 5. 13 Gráfico de barras de emoções França vs. Bélgica 48
Figura 5. 14 Test&Score França vs. Bélgica 48
Figura 5. 15 Gráfico de barras de emoções Inglaterra vs. Croácia 49
Figura 5. 16 Test&Score Inglaterra vs. Croácia 49
Figura 5. 17 Gráfico de barras de emoções Inglaterra vs. Bélgica 49
Figura 5. 18 Test&Score Inglaterra vs. Bélgica 50
Figura 5. 19 Gráfico de barras de emoções comparativo de todos os jogos analisados 50
Figura 5. 20 Gráfico de barras de emoções relativos ao jogador Cristiano Ronaldo e a Juventus 51
Figura 5. 21 Test&Score Cristiano Ronaldo - Juventus 51
Figura 5. 22 Gráfico de barras de emoções relativos ao jogador Cristiano Ronaldo e o Real Madrid 52
Figura 5. 23 Test&Score Cristiano Ronaldo – Real Madrid 52
Figura 5. 24 Gráfico de barras de emoções comparativo dos datasets que relacionam o Cristiano Ronaldo com a Juventus e Real Madrid 52

Lista de Tabelas

Tabela 3. 1 Tabela comparativa de *softwares*..... 32

Lista de Acrónimos

UBI	Universidade da Beira Interior
PLN	Processamento de Linguagem Natural
FIFA	*Fédération Internationale de Football Association*

Folha em branco

Capítulo 1

1. Introdução

A presente dissertação insere-se no contexto da análise de sentimentos baseadas na extração de informação de redes sociais. Com o objetivo de validar o modelo proposto, dois casos de estudo foram considerados:

I. Análise do evento Mundial FIFA 2018 de futebol que se realizou na Rússia;
II. Análise sobre a transferência do jogador Cristiano Ronaldo do Real Madrid CF para a Juventus FC e sobre as repercussões que isso teve ao nível das redes sociais.

A análise foi feita na rede social *Twitter* que oferece um serviço de *microblogging*, onde cada utilizador pode publicar *tweets* sobre os mais variadíssimos temas. Neste caso foram apenas considerados os *tweets* relacionados com os casos de estudo propostos.
Em primeiro lugar foi realizada uma pesquisa intensa sobre que rede social analisar e chegou-se à conclusão que o Twitter seria e melhor escolha devido aos seus tweets serem curtos (o *Twitter* tem um limite de 280 caracteres), de fácil análise e pelo facto de esta rede social possuir uma elevada taxa de utilizadores.
De seguida procedeu-se à escolha da linguagem de programação a utilizar para o download dos tweets, tenho a escolha recaído sobre a linguagem *Python*, pelo facto de ser uma linguagem de fácil aprendizagem, muito intuitiva e também por ser uma das mais utilizadas no contexto de *machine learning*.
Após a escolha da rede social e da linguagem de programação, chegou a altura de escolher a *framework* para ser realizada a análise dos dados do *Twitter*. Depois de muita pesquisa foi escolhida a *framework Orange 3* porque esta ferramenta é gratuita e tem os mais variadíssimos recursos para a mineração de dados e também para a mineração de texto que é o núcleo desta dissertação.

1.1 Objetivo

O objetivo desta dissertação é de saber o que é dito nas redes sociais sobre os casos de estudo propostos e, através disso, fazer uma análise de sentimentos dos *tweets* dos utilizadores. Pretende-se assim conhecer as diferentes emoções que são expressas nos comentários e analisar detalhadamente as emoções expressas sobre as diferentes seleções e sobre os jogos das mesmas.
Esta análise, no caso de estudo do Mundial FIFA 2018, foi feita com base nas seguintes seleções nacionais: França, Croácia, Bélgica, Inglaterra e Portugal. Esta escolha deve-se ao facto de terem sido as primeiras quatro seleções que chegaram às últimas fases da

competição. A escolha de Portugal deve-se ao facto de ser uma das seleções favoritas (por ter ganho o Europeu de 2016) e ter ficado pelo caminho nos "oitavos-de-final".
Já no segundo caso de estudo relacionado com o jogador Cristiano Ronaldo, foi feita uma análise que incidiu sobre a relação do jogador com as duas equipas em questão, que são o Real Madrid CF e Juventus FC.

1.2 Motivação

É cada vez mais relevante saber o que é dito nas redes sociais sobre os mais variados temas e, apesar de haver bastantes trabalhos de pesquisa nesta área da análise de sentimentos, relativamente aos casos de estudo propostos o número é reduzido ou praticamente inexistente. É importante conhecer as emoções que os adeptos afetos, tanto às seleções em causa no mundial de futebol, como ao jogador Cristiano Ronaldo e equipas correspondentes sentem. Isto porque os dois casos de estudo em causa envolvem um número bastante elevado de pessoas de países em todo o mundo.

1.3 Questões de Pesquisa

Nesta dissertação vai tentar responder-se às seguintes questões, organizadas por dois casos de estudo:

i) **Caso de estudo "Mundial FIFA 2018":**

- Qual a equipa cujos comentários dos adeptos têm uma polaridade mais positiva?
- Como é que os jogos do *FIFA World Cup 2018* afetaram os comentários dos utilizadores e que tipo de emoções estes expressaram?

ii) **Caso de estudo "efeito Cristiano Ronaldo":**

- Como é que a transferência do jogador Cristiano Ronaldo do Real Madrid para a Juventus afetou os comentários dos utilizadores afetos aos dois clubes?

1.4 Enquadramento

A Web é um dos principais meios de comunicação, onde existem massivas quantidades de dados, e assim sendo, muitas empresas estão interessadas em saber o que é dito acerca delas para assim poderem gerir a sua reputação online. [1]

É comum um cliente que utiliza um certo produto ou usufrui de um certo serviço de uma empresa, deixe a sua opinião em formato textual, ou em redes sociais ou até no Web Site da empresa.
O que o cliente escreve sobre o produto ou serviço tem uma certa polaridade, que pode ser positiva, negativa ou neutra. [2] A polaridade é a orientação do sentimento expresso numa frase.
Quando a quantidade de opiniões por parte dos clientes é demasiado alta, analisar as mesmas manualmente torna-se praticamente impossível, sendo necessário realizar este processo de forma automática. É precisamente aqui que entra o conceito de Mineração de Texto e Análise de Sentimentos, que por sua vez são uma área da Aprendizagem Automática. A Mineração de Texto usa também técnicas de Processamento de Linguagem Natural (PLN) [3] que têm como objetivo melhorar a compreensão da linguagem humana através de técnicas para processar textos rapidamente [4]. Faz uso de técnicas como por exemplo *tokenization* e remoção de *stop-words* (como explicado na secção 2.3.1).

1.5 Estrutura da dissertação

A presente dissertação está estruturada da seguinte forma:

- O primeiro capítulo é o capítulo introdutório onde se aborda e contextualiza o tema e onde se definem as questões de pesquisa que guiaram esta dissertação;
- O segundo capítulo trata-se do Estado-da-Arte da disciplina da Análise de Sentimentos, Mineração de Texto e Mineração de Dados.
- De seguida, no terceiro capítulo são abordadas e discutidas as tecnologias implementadas;
- No quarto capítulo são apresentados todos os passos seguidos na fase de implementação;
- O quinto capítulo está inserido na discussão dos resultados obtidos com a implementação realizada e, é também onde se respondem às questões de pesquisa propostas.
- Por fim, no sexto capítulo é feita a conclusão do processo de realização desta dissertação e possíveis trabalhos futuros.

Capítulo 2

2. Estado da Arte

Neste capítulo será abordado o Estado da Arte da Análise de Sentimentos, e de seguida será também abordado o tema de Mineração de Texto (*Text Mining*) e Mineração de Dados *(Data Mining*). Irá ser abordada a importância da Análise de Sentimentos, as suas vantagens, os seus diferentes tipos e as suas aplicações práticas.

2.1 Análise de Sentimentos

A análise de sentimentos (AS) é um tema que tem ganho uma grande importância devido a poder ser aplicável às mais variadíssimas questões, que podem ser a monitorização de redes sociais, de marcas, serviço ao cliente, análise de produtos, ou até o estudo e análise do mercado [1].
A AS pode ser entendida como o processo automatizado de extrair uma opinião sobre um determinado assunto (marca, evento, etc.) de um texto em linguagem natural [1], também conhecida como mineração de opinião, que é um tema dentro da área do Processamento de Linguagem Natural (PLN). O propósito é construir sistemas automatizados que possam identificar e extrair opiniões em texto não-estruturado.

2.1.1 Importância da Análise de Sentimentos

A maior parte da informação contida na Web, tanto em e-mails, *chats*, redes sociais, artigos, documentos, encontra-se em formato não-estruturado, tornando-se assim dispendioso analisá-la manualmente. A análise de sentimentos possibilita tanto às empresas, analistas, ou investigadores, uma análise de forma automática deste tipo de informação.

2.1.2 Vantagens da Análise de Sentimentos

Algumas das vantagens mais importantes da Análise de Sentimentos [1] são:

- **Escalabilidade**: consiste em fazer de forma automática o que seria impossível fazer manualmente como, por exemplo, analisar milhares de tweets;
- **Análise em tempo-real:** a análise pode ser usada para poder analisar cenários específicos em tempo-real, isto porque pode ser importante identificar estas situações e preveni-las;
- **Critérios consistentes:** definir critérios para a análise de sentimentos é uma tarefa subjetiva, já que pode ser influenciada por experiências pessoais, entre outras coisas.

Para evitar este tipo de obstáculos é definido um sistema de análise de sentimentos centralizado em que é aplicado o mesmo critério para todos os casos.

2.1.3 Tipos de Análise de Sentimentos

Existem diversos tipos de análise de sentimentos e neste tópico vão ser abordados alguns dos mais importantes [1]:

- **Deteção de Emoções**: este tipo de análise tem o objetivo de, como o próprio nome indica, detetar emoções que podem ser, por exemplo, felicidade, tristeza, repugnância, medo, surpresa e raiva, fazendo uso de um dicionário de palavras (*lexicon*) em que cada uma tem atribuída um tipo de emoção;

- **Baseada em Aspetos**: este tipo de análise permite fazer a ligação de um sentimento relativamente a um aspeto ou entidade;

- **AS em diferentes idiomas:** pode ser uma tarefa bastante difícil, já que maior parte dos sistemas estão preparados apenas para fazer análise em inglês. Fazer a análise em idiomas diferentes requer um dicionário de sentimentos, onde cada palavra tem um sentimento associado.

2.1.4 Aplicações práticas

Existem diversas hipóteses de aplicação da AS. As seguintes são algumas das mais comuns:

Sentimento sobre um determinado produto/serviço:

- A análise de sentimentos tem a vantagem de dar a uma empresa a possibilidade de saber o que os consumidores acham acerca dos seus produtos, identificando assim ideias e sugestões que podem melhorar os seus produtos de forma a satisfazer os consumidores;
- Análise por parte de empresas mais diretamente ligadas ao estudo do mercado, na medida em que compara as análises de sentimentos do seu produto ou serviço com as da concorrência e estuda as tendências do mercado;
- Pode ser útil também gerar relatórios com uma certa frequência para ir tendo feedback sobre o produto ou serviço.

Monitorização de redes sociais:

- A aplicabilidade da AS dá uma perspetiva sobre um dado acontecimento, evento, entre outros [2]. É possível assim, por exemplo, numa campanha eleitoral saber, por

parte de quem faz a análise, quais os candidatos com maior aceitação por parte do público alvo.

2.1.5 Métodos de Classificação de Sentimentos

Existem muitos algoritmos de AS, que podem ser divididos em três classes [1]:

- métodos baseados em regras;
- métodos automáticos;
- métodos híbridos.

- **Métodos baseados em regras:**

    - Como o próprio nome indica, estes métodos baseiam-se em regras que podem ser escritas para identificar a subjetividade, polaridade ou até o assunto em questão;
    - Estes métodos podem usar várias técnicas de Processamento de Linguagem Natural (PLN), como por exemplo, *stemming*, *tokenization*, etc;
    - Podem recorrer também a *lexicons,* que consistem em dicionários de palavras, em que cada palavra tem um sentimento atribuído.

- **Métodos Automáticos**
    - Contrariamente aos métodos baseados em regras, estes fazem uso de técnicas de *machine learning* para fazer a AS;
    - Modelam o problema como se fosse um problema de classificação, em que recebem um pedaço de texto como *input* e devolve a polaridade correspondente como *output*.

- **Métodos Híbridos**
    - Estes métodos combinam as abordagens dos métodos baseados em regras e dos métodos automáticos.

2.1.6 Avaliação do Sentimento

Após ter sido feita a AS no conjunto de dados, é necessário recorrer a algoritmos de avaliação. Estes vão calcular a precisão (*precision*), cobertura (*recall*), a medida F (*F-Measure*) e a taxa de acerto (accuracy) [3]. Os cálculos são feitos como está explicado de seguida.

$$Precisão = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Cobertura = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Medida\ F = 2 * \frac{Precisão * Cobertura}{Precisão + Cobertura}$$

$$Taxa\ de\ acerto = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Negative + False\ Positives}$$

True Positives corresponde aos valores que foram previstos positivos e são de facto positivos;
True Negatives corresponde aos valores previstos como negativos e são de facto negativos;
False Positives corresponde aos valores que são previstos como positivos, mas são de facto negativos;
False Negatives corresponde aos valores que são previstos como negativos, mas são de facto positivos.

A **precisão** corresponde ao número de valores que foram corretamente classificados. A **cobertura** corresponde ao número total de valores classificados como positivos.
A **Medida F** combina os valores de precisão e cobertura para obter uma pontuação (*score*) e pode servir para comparar algoritmos de classificação para ver o que tem melhor pontuação.
A **Taxa de Acerto** é a medida mais intuitiva e faz o cálculo das previsões corretas no número total de valores.

2.2 Análise de Sentimentos no Twitter

A rede social *Twitter* possibilita uma rápida distribuição de informação dentro de uma grande população de utilizadores, sendo altamente eficaz para disseminar notícias em tempo real. [4] É também uma das maiores redes sociais do mundo que oferece serviços de *microblogging*. Possibilita, além disso, a publicação de mensagens, que são conhecidas por *Tweets*, de até 280 caracteres.
Em março de 2018 o *Twitter* tinha cerca de 330 milhões de utilizadores ativos por mês e eram publicados cerca de 500 milhões de *Tweets* diariamente. [5] Devido a este número elevados de utilizadores é de extrema utilidade fazer uma análise de sentimentos nesta rede social para se descobrir o que é dito por parte dos utilizadores em relação a uma certa marca, empresa ou evento.

2.3 Mineração de Texto

A temática da Mineração de Texto tem vindo a tornar-se uma importante área de pesquisa, já que consiste na descoberta de informação previamente desconhecida em grandes quantidades

de texto não-estruturado. Esta descoberta é conseguida através de processos automáticos de extração de informação.

A motivação da Mineração de Texto vem do facto de ser bastante difícil pôr de parte informação que não seja relevante, para extrair apenas a informação relevante pretendida.

A Mineração de Texto, também conhecida por Descoberta de Conhecimento em Texto (*Knowledge-Discovery in Text*), tem como objetivo principal extrair conceitos e relações entre esses conceitos usando técnicas de Processamento de Linguagem Natural (PLN) [6].

De seguida iremos explicar cada uma das etapas da Mineração de Texto [7].

2.3.1 Etapas da Mineração de Texto

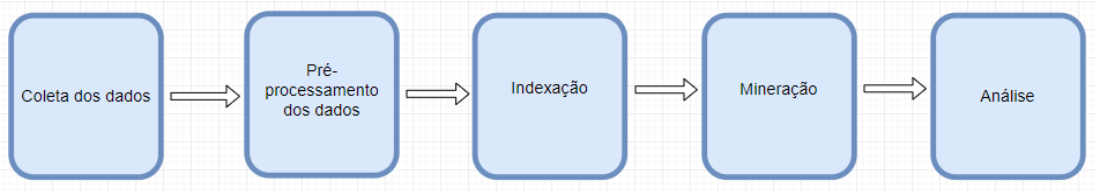


Figura 2. 1 Etapas da Mineração de Texto

- **Coleta de dados** de diferentes fontes para se obter um *dataset*;
- **Pré-processamento** dos dados. Este passo faz uso de técnicas de PLN como por exemplo: remoção de *stop-words*, ou seja, retirar as palavras que não acrescentem qualquer tipo de valor para o *dataset*; *Tokenization*, que significa uma divisão do texto com o objetivo de obter entidades significativas e com mais valor para análise; Extração de características (*Feature Extraction)*, que é o processo de extração das entidades do texto, como por exemplo a frequência das palavras, os conceitos, etc.;
- **Indexação** dos termos para se obter um melhor acesso para efeitos de consulta;
- **Mineração**: é nesta fase que os dados são explorados para se obter novo conhecimento, como por exemplo, fazendo uma comparação dos termos com um dicionários (*lexicons)* e relacionando esses termos;
- **Análise**: última fase da mineração de texto, onde se avalia e visualiza os resultados obtidos, consoante as questões a que se quer responder.

2.4 Mineração de Texto VS. Mineração de Dados

A mineração de texto, que é uma subárea da mineração de dados, vem responder aos mais diversos problemas, tanto ao nível empresarial como académico, de análise textual.

Como a maior parte da informação disponível se encontra na forma semiestruturada ou não-estruturada, torna-se necessário recorrer ao processo de mineração de texto.

De seguida apresentam-se alguns pontos de comparação entre os processos de mineração de texto e mineração de dados [8]:

- A mineração de texto aplica as mesmas funções analíticas que a mineração de dados, mas neste caso, aplicadas ao domínio textual, baseadas em análise textual sofisticada que extrai informação de documentos de texto;
- A mineração de dados inclui uma parte limitada dos dados porque a maior parte dos mesmos não se encontra estruturada;
- A mineração de texto lida com informação estruturada ou semiestruturada;

Portanto, para melhorar o processo de extração de padrões úteis em formato textual, torna-se imperativo a escolha do processo de Mineração de Texto.

2.5 Revisão da Literatura

Em 2015, Sarah E. Shukri, et al. [9] publicaram um caso de estudo de análise de sentimentos no *Twitter* sobre a indústria automóvel. Como a indústria automóvel é um dos maiores setores económicos e devido ao facto de a concorrência ser elevada, as companhias deste setor estão a ir de encontro às redes sociais para chegar aos seus clientes ou futuros clientes. Este estudo usou a rede social *Twitter* como fonte de dados e extraiu opiniões de utilizadores em relação a três companhias: Mercedes, Audi e BMW. Estes *tweets* foram extraídos usando técnicas de mineração de texto e de seguida classificados usando Análise de Sentimentos.

Em 2015, Elif Uysal, et al. [10] escreveram um artigo sobre o caso de estudo das eleições da Turquia em 2015. Usaram também, a rede social *Twitter* como fonte de dados. Foi criado um dicionário em língua turca com palavras-chave com um sentimento específico. Foi então usado este dicionário para detetar a polaridade dos *tweets* com determinadas palavras-chave relacionadas com política, imediatamente antes das eleições em 2015. Os t*weets* foram recolhidos de acordo com a sua relação com as seguintes categorias: líderes políticos, ideologias e partidos políticos. Os resultados da classificação dos *tweets* foram posteriormente analisados e comparados com a resultado das eleições.

Em 2014, Wu He, et al. [11], publicaram um estudo em que analisam conteúdo da rede social *Twitter* e *Facebook*, sobre três das maiores cadeias de *fast-food*, mais propriamente sobre *pizza*. As cadeias estudadas são: *Pizza Hut*, *Domino's Pizza* e *Papa John's Pizza*. Foram recolhidos dados das duas plataformas, como por exemplo o número de fãs/seguidores, número de publicações feitas, número de comentários, a frequência com que foram feitas essas publicações e o tempo de resposta. Nos resultados, no caso do *Twitter*, foram descobertos os seguintes temas: encomendas e entregas; qualidade da *pizza*; *feedback* dos consumidores; *tweets* casuais; *tweets* de *marketing*; No caso da rede social *Facebook* foram encontrados os seguintes temas: publicação de imagens, de perguntas para obter resposta por parte dos utilizadores, atividades da companhia e da comunidade e por fim, publicação de informações e promoções. Os resultados foram analisados com o intuito de se fazer uma

comparação das três cadeias nas duas plataformas. Estes resultados mostraram que estas cadeias de *fast-food* estão todas a fazer esforços para manterem uma reputação respeitável nas redes sociais com o objetivo de melhorar a satisfação dos clientes.

Capítulo 3

3. Tecnologias implementadas

Neste capítulo vão ser abordadas, não só as diferentes tecnologias utilizadas para a realização desta dissertação, mas também a justificação da escolha das mesmas, comparativamente com alternativas.

3.1 Linguagem de programação para obter os *datasets*

Nesta fase, uma das primeiras a ser realizadas, foram estudadas várias linguagens de programação que poderiam ter sido usadas no âmbito desta dissertação. Teria que ser uma linguagem com possibilidade de ligação à rede social *Twitter*, para de facto, extrair os dados, mas também teria que ser intuitiva e versátil.

3.1.1 Java

No início da realização deste projeto estudou-se a hipótese de implementar esta linguagem de programação, devido a possuir possibilidade de fazer a ligação entre o programa em si com a rede social *Twitter,* através de bibliotecas de *Java* e existir bastante informação disponível acerca disso.

3.1.2 Python

A escolha acabou por recair na linguagem de programação *Python*. Uma das razões pelas quais foi feita esta escolha foi o facto de ser das mais usadas em contexto de *machine learning* e mineração de texto.
A linguagem *Python* é também uma das mais usadas em companhias de programação [12] e uma das mais versáteis já que pode ser aplicada nas mais diversas áreas, tais como a área de *gaming,* ferramentas *web service*, prototipagem, etc.
Possui, além disso, bibliotecas em abundância, o que representa uma grande vantagem para esta linguagem. Estas bibliotecas possibilitam a integração do código *Python* nas mais diversas áreas, sendo uma delas a *Data Science* (Ciência dos Dados). Exemplos destas bibliotecas são o *NumPy, SciPy, pandas, etc*. Uma das bibliotecas mais importantes para esta dissertação é o *tweepy*, que possibilita a interação do código *Python* com o *Twitter*.

3.2 *Framework* para análise de sentimentos

Posteriormente à escolha da linguagem de programação, procedeu-se à escolha da *framework* a usar para se realizar a análise de sentimentos e classificação e avaliação dessa análise. Esta

framework teria que ser *open-source* e teria que responder às diferentes exigências abordadas, ou seja, teria que possuir ferramentas de mineração de texto, entre as quais a análise de sentimentos e teria que ter a possibilidade de implementar algoritmos de classificação.

De seguida irá ser apresentada uma tabela comparativa com as ferramentas que poderiam ter sido utilizadas [13], com algumas das suas funcionalidades mais importantes:

Orange 3	*RapidMiner*	*KNIME*
• Ferramenta *open source;* • Permite visualização de dados interativa; • Programação visual; • Orientado tanto para iniciantes como para profissionais; • Executa análise de dados simples e complexos; • Possui ferramentas para análise de sentimentos em redes sociais e não só.	• *Freeware* com versões a pagar; • Múltiplas funções de *machine learning;* • Visões gerais ao nível estatístico; • Cria *datasets* ótimos para análise preditiva; • Limpeza de dados para algoritmos; • Permite análise de sentimentos.	• *Software* grátis; • Várias funcionalidades matemáticas e estatísticas; • Algoritmos de *machine learning*; • Visualização de dados interativa; • Possui ferramentas para análise de sentimentos em redes sociais.

Tabela 3. 1 Tabela comparativa de *softwares*

3.2.1 Orange 3

A tecnologia que acabou por escolhida foi a *Orange 3*, porque tem a vantagem de ser gratuita sem períodos experimentais e conseguir fazer o mesmo tipo de análise que outras ferramentas.

É uma *framework open-source* para mineração de dados, mineração de texto, *machine-learning*, visualização de dados, entre outras funcionalidades. É uma ferramenta que possibilita a análise de dados de uma forma interativa.

É *user-friendly*, poderosa, rápida e possui uma programação visual versátil para visualização de dados e análise. Foi desenvolvida em *C++* e *Python.*

Possui os mais variadíssimos *add-ons* que vão desde a bioinformática, análise de imagens, até à mineração de texto, que é essencial para esta dissertação. Dentro destes *add-ons*, a *Orange* possui uma coleção bastante completa de *widgets*. *Widgets* estes que possibilitaram a análise de sentimentos dos dados em causa, o pré-processamento do texto, algoritmos de classificação, entre outros.

Capítulo 4

4. Implementação

Neste capítulo irá ser abordada a fase de análise de requisitos desta dissertação. Vai ser discutida a linguagem *Python* e como foi usada para obter os *datasets*, e a *framework Orange 3*, e como esta foi usada para implementar um sistema de análise de sentimentos.
Em baixo é apresentado um esquema ilustrativo das etapas que foram realizadas ao longo deste projeto.

Figura 4. 1 Etapas da Implementação

A primeira etapa foi realizada através de um programa em *Python* e, depois disso, as restantes etapas foram elaboradas através da *framework Orange 3*.

4.1 Python

A linguagem de programação usada foi a linguagem *Python*, porque, como foi dito anteriormente, é uma das mais usadas e é também das mais versáteis.
Foi usada esta linguagem para obter os *datasets* pretendidos, que são os *tweets* relacionados com a os dois casos de estudo em questão.
Foi usado o *IDE PyCharm* para o desenvolvimento deste código *Python*.

4.1.1 Extensões *Python*

Na linguagem *Python*, como em muitas outras, temos a possibilidade de usar as mais variadíssimas extensões (*packages*) que nos permitem fazer ponte de ligação do código *Python* com muitas outras plataformas.
Ao longo desta dissertação foram usados as seguintes extensões:

- ***Textblob*** – que faz uso de técnicas de mineração de texto;
- ***Tweepy*** – que faz a ligação entre o código *python* e o *Twitter*;
- ***CSV*** – que permite manipular dados no formato *.CSV*.

4.2 Implementação *Python*

Primeiramente foi criada uma aplicação na rede social *Twitter*, para se poderem obter as chaves de acesso necessárias para fazer *download* dos *tweets* pretendidos. De seguida, inseriram-se as chaves de acesso expostas em cima, no código *Python*.
Para se poder fazer *download* dos *datasets* fez-se a inserção das chaves de acesso do *Twitter*, de seguida criou-se um ficheiro de formato *.CSV* para guardar esse *dataset* e por fim inseriu-se a *query* pretendida, que no âmbito desta dissertação foi sobre as seleções escolhidas e dos jogos entre elas.
Estas *queries* correspondem aos parâmetros escolhidos, que definem que o limite dos *tweets* vai até 10000 *tweets* por *query* e que estes têm que estar no idioma inglês.

As *queries*, para o caso de estudo "Mundial FIFA 2018", foram as seguintes:

- *"France National Team World Cup 2018"*
- *"Croatia National Team World Cup 2018"*
- *"Belgium National Team World Cup 2018"*
- *"England National Team World Cup 2018"*
- *"Portugal National Team World Cup 2018"*
- *"France Croatia match World Cup 2018"*
- *"France Belgium match World Cup 2018"*
- *"Croatia England match World Cup 2018"*
- *"England Belgium match World Cup 2018"*

Já as *queries* relativas ao segundo caso de estudo "efeito Cristiano Ronaldo" foram as seguintes:

- *"Cristiano Ronaldo Real Madrid"*
- *"Cristiano Ronaldo Juventus"*

De seguida, fez-se uma otimização dos *tweets*, para que não existissem *tweets* duplicados, como é demonstrado no seguinte pedaço de código *Python:*

```
if (not tweet.retweeted) and ('RT @' not in tweet.text):
print (tweet.created_at, tweet.text)
```

4.3 Orange 3

A *framework* usada, como já referido anteriormente, foi o *Orange 3*. Foi escolhida esta plataforma devido a ser *open-source*, estar preparada para fazer análises de *Text Mining* e *Data Mining* e ter uma coleção vasta de funcionalidades que permitem fazer as análise e tratamento dos dados pretendidos.

4.3.1 *Add-ons* e funcionalidades de mineração de texto

Nesta plataforma podemos usufruir dos mais diversos *Add-ons*, que vão desde ferramentas de mineração de texto, análise de imagens, bioinformática, etc. O a*dd-on* usado no âmbito desta dissertação foi o da mineração de texto, já que possui as mais diversas funcionalidades para análise de texto.
As funcionalidades usadas no âmbito da Mineração de Texto foram os seguintes: *Sentiment Analysis, Tweet Profiler, Corpus, Word Cloud, Preprocess Text*.

4.3.1.1 *Sentiment Analysis*

Como o próprio nome indica, esta funcionalidade serve para fazer Análise de Sentimentos, usando o método *VADER* [14]. Este método faz uso de um *lexicon*, que, como já foi dito, consiste num dicionário de palavras com um sentimento atribuído a cada uma delas. Dicionário este que é elaborado usando métodos qualitativos e quantitativos.

4.3.1.2 *Tweet Profiler*

O objetivo do *Tweet Profiler* é o de fazer uma análise de sentimentos e devolver o tipo de sentimento por classes que podem ser felicidade, tristeza, repugnância, medo, surpresa e raiva.

4.3.1.3 *Distributions*

O propósito desta funcionalidade é o de demonstrar, em forma de gráfico, a distribuição dos valores discretos ou contínuos de uma determinada variável.

4.3.1.4 *Corpus*

Esta funcionalidade é a mais simples de todos, porque faz apenas a inserção do *Corpus* obtido com o programa *Python*, e insere-o no *workflow* do *Orange 3*.

4.3.1.5 *Word Cloud*

Esta funcionalidade mostra numa tabela de visualização as palavras mais frequentemente usadas com a frequência correspondente, num dado *corpus*.

4.3.1.6 *Preprocess Text*

Esta funcionalidade faz uso de técnicas da área do Processamento de Linguagem Natural para pré-processar o texto.
O que faz concretamente é dividir o texto em unidades mais pequenas, chamadas de *tokens*, que neste caso faz uma divisão segundo um modelo já feito sobre o *Twitter*, que faz com que se mantenham as *hashtags* e outros símbolos especiais. De seguida faz uma filtragem do texto, que consiste em remover *stopwords*, que são palavras que podem aparecer muitas vezes não acrescentando valor nenhum à classificação e análise de sentimentos, como é o caso das palavras "I", "me", "the", "you", etc.
Faz também uma transformação dos *tweets* em causa, ao remover a acentuação, *urls,* etc.

4.3.1.7 *Topic Modelling*

Nesta funcionalidade a tarefa realizada é a modelação de tópicos, e como o próprio nome indica, faz uma espécie de divisão por tópicos, as palavras mais utilizadas juntas.

4.3.2 Funcionalidades de visualização e manipulação de dados

As funcionalidades mencionadas anteriormente faziam parte do *add-on* da mineração de texto. No entanto, foram usados muitos outros. Estas funcionalidades faziam as mais variadíssimas tarefas possibilitando a realização desta dissertação, como foi o caso da classificação da análise de sentimentos e a visualização e manipulação dos dados.
As funcionalidades seguintes referem-se à visualização e manipulação de dados.

4.3.2.1 *Data Table*

Esta funcionalidade representa uma tabela de dados, que permite a sua visualização.

4.3.2.2 *Select Columns*

A função desta funcionalidade é a de selecionar colunas de um dado *input* de dados, para devolver apenas as colunas selecionada como *output*.

4.3.2.3 *Data Sampler*

O *data sampler* tem como propósito dividir os dados de *input*, para, neste caso, serem atribuídos dados para teste e dados para treino. Por padrão, 70% dos dados são atribuídos para propósitos de treino e os restantes 30% são atribuídos para propósitos de teste.

Estes dados já separados são atribuídos para dois caminhos diferentes, os de treino seguem para os algoritmos de classificação, que irão ser referidos mais adiante, enquanto que os de teste são logo encaminhados para uma tabela de "*test & score*" (teste e resultado).

4.3.2.4 *Discretize*

A funcionalidade *discretize* tem como objetivo discretear os atributos contínuos através de um método específico. O método escolhido foi o de *equal-width* (igual-largura) que consiste em dividir igualmente o intervalo entre o maior e o menor valor observado.
Esta funcionalidade no caso presente, tem como *input* o *data sampler*, para devolver de seguida por um lado, para os algoritmos de classificação usados, e por outro, para ligar diretamente para a tabela de *test & score*.

4.3.2.5 *Scatter Plot*

Esta funcionalidade proporciona um gráfico de dispersão bidimensional de visualização tanto para atributos contínuos como para atributos discretos. Os dados são apresentados como pontos do gráfico cada um tendo um valor no eixo do X e no eixo do Y.

4.3.3 Funcionalidades de classificação de dados

As funcionalidades abordadas de seguida fazem parte da categoria de classificação do *Orange 3*, e o seu propósito é classificar os dados provenientes da Análise de Sentimentos realizada. Os métodos seguintes pertencem à área de aprendizagem supervisionada.

4.3.3.1 *Neural Network*

Esta funcionalidade é relativa ao algoritmo de Redes Neuronais, mais concretamente ao algoritmo de perceptrão multicamada. Estas redes são compostas por nodos, que não são mais que elementos de processamento., interligados entre si. Este tipo de algoritmo é responsável pelo processo de classificação e consiste na emulação dos neurónios humanos e da maneira como estes reconhecem padrões.
Este algoritmo possui várias camadas, cada uma com vários neurónios e com cada neurónio a ter um certo peso de conexão atribuído. Cada camada tem variáveis de entrada e variáveis de saída para a camada seguinte.

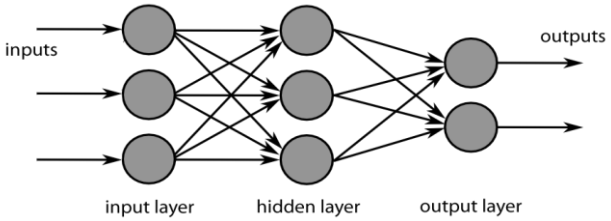


Figura 4. 2 Rede Neuronal

A aprendizagem de uma rede deste tipo consiste em manipular os pesos da rede para minimizar o erro entre o *output* da rede e o objetivo que é pretendido. [15]

Este tipo de algoritmos é bastante eficiente para classificar padrões não-lineares como é o caso dos objetos de estudo nesta dissertação.

Depois de ajustados os pesos de cada neurónio, este algoritmo tem que passar por uma função de ativação que neste caso é a função *ReLu*. Uma vantagem desta função de ativação é que esta não ativa todos os neurónios ao mesmo tempo [16], ou seja, se o *input* de um neurónio for negativo, este não será ativado. Esta característica faz com que esta função seja menos dispendiosa em termos computacionais.

Esta função é também, das redes neuronais mais usadas atualmente [15] em *deep learning*.

4.3.3.2 *Naive-Bayes*

O classificador *Naive-Bayes* é um dos classificadores mais simples, devido à matemática simples envolvida [17]. Por ser um algoritmo simples, muitas vezes consegue obter resultados melhores que outros classificadores.

É um algoritmo *Naive* (ingénuo) porque não considera que haja qualquer tipo de relação entre as variáveis que possui.

É um classificador baseado no teorema de *Bayes:*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) é a probabilidade de o evento "A" acontecer dado que o evento "B" acontece.

P(B|A)P(A) é a probabilidade de o evento "B" acontecer dado que o evento "A" acontece, multiplicado pela probabilidade de acontecer o evento "A". Isto é tudo dividido de seguida pela probabilidade de o evento "B" acontecer.

4.3.3.3 *Knn*

Este algoritmo de classificação é baseado na proximidade dos vizinhos, mais concretamente na proximidade do vizinho mais próximo (*Nearest Neighbor*). Esta técnica tem como objetivo o reconhecimento de padrões através dos vizinhos que se encontrem mais próximos [18].

O algoritmo *kNN* faz a classificação de uma dada instância, calculando a distância entre essa instância e os dados de treino. Isto é conseguido através de uma função de cálculo de distância, que neste caso, é a distância Euclidiana. Esta função é demonstrada de seguida:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Esta função classifica a instância dependendo da proximidade para com as outras instâncias (*k)* de uma determinada classe, classificando-a com essa classe.

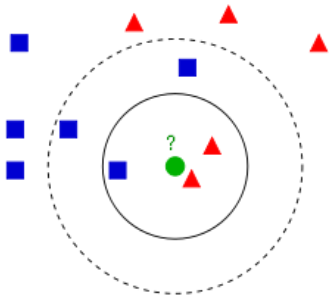


Figura 4. 3 *kNN*

4.3.3.4 *Logistic Regression*

O algoritmo de regressão logística (*Logistic Regression)* é também um dos algoritmos mais usados dentro da disciplina de *machine learning*, para classificação binária.

Este algoritmo mede a relação entre a variável dependente (aquela que pretendemos prever) e as restantes variáveis dependentes, estimando as probabilidades usando a sua função logística.

Este modelo está relacionado com o modelo de regressão linear (*Linear Regression)*, mas usa uma função logística inversa para transformar o valor do *output* num valor entre zero e um [19], valor este que pode ser interpretado como uma variável.

$$P(X) = \frac{e^{(\beta 0 + \beta 1 * X)}}{(1 + e^{(\beta 0 + \beta 1 * X)})}$$

É usado para determinar o *output* quando há uma ou mais variáveis independentes.

Algumas das vantagens deste algoritmo é que é bastante eficiente e não consome muitos recursos computacionais.

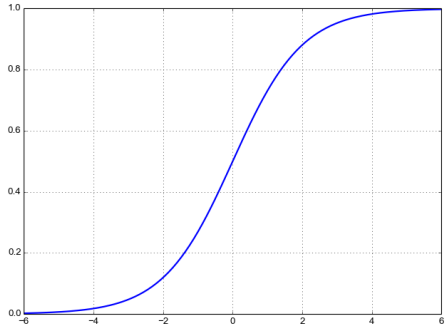


Figura 4. 4 Regressão Logística

4.3.3.5 *Support Vector Machine (SVM)*

O algoritmo *SVM* é um dos algoritmos mais poderosos para categorização textual. Combina métodos estatísticos com métodos de *machine learning* para gerar funções de mapeamento de *input-output*. Tem como *input* um vetor e como *output* zero ou um (positivo ou negativo). Este algoritmo representa as instâncias (*features*) como pontos no espaço, mapeados para que os exemplos das diferentes categorias sejam separados por uma margem tão ampla quanto possível. Faz parte da categoria de *supervised machine learning algorithms*, e tanto pode ser usado para problemas de classificação como para problemas de regressão (*regression*) [20].

A classificação é feita encontrando o hiperplano que diferencia as duas classes de forma eficiente, como é demonstrado na seguinte figura.

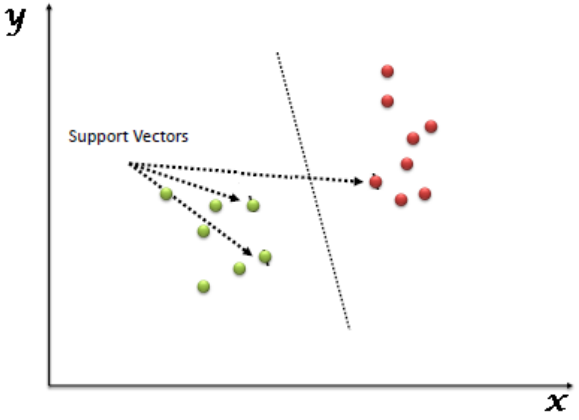


Figura 4. 5 Support Vector Machine

4.3.4 Funcionalidades de avaliação

Até agora fez-se a análise de sentimentos e a sua classificação de polaridade. Chegou a altura de avaliar os resultados obtidos.

4.3.4.1 *Test & Score*

Esta funcionalidade avalia os resultados da classificação feita anteriormente. Faz o cálculo da Precisão (*precision)*, Cobertura (*recall),* Taxa de Acerto (*classification accuracy*) e da Medida F (*F Measure)*. Estes resultados vão mostrar se os algoritmos de classificação usados tiveram um bom desempenho ou não.

4.4 Implementação *Orange 3*

Fazendo uso do modelo proposto na secção 4, fez-se a junção de todas as funcionalidades usadas para se poder obter o *workflow* para a análise de sentimentos. De seguida, explica-se detalhadamente, como se juntaram essas funcionalidades:

1. Em primeiro lugar, fez-se a inserção do *dataset*, adquirido com a implementação do código *python*, na funcionalidade *Corpus;*
2. De seguida foi feito um pré-processamento desses documentos, para se poder obter um *dataset* mais "limpo" e preparado para análise;
3. Ligados à funcionalidade de pré-processamento, encontram-se dois outras funcionalidades que são o *Topic Modelling*, que organizam por tópicos as palavras mais usadas juntas, e a funcionalidade *Word Cloud* (nuvem de palavras), que demonstra um gráfico com as palavras mais usadas e a sua frequência;
4. Procedeu-se então à análise de sentimentos propriamente dita, que consistiu na atribuição de uma certa polaridade a cada *tweet*;
5. Ligado à funcionalidade de análise de sentimentos encontra-se a funcionalidade *Tweet Profiler,* que como foi dito anteriormente, atribui um sentimento específico a cada *tweet*, quando complementado com a funcionalidade *Distibutions;*
6. Após a análise de sentimentos, fez-se uma seleção das colunas para podermos analisar apenas as variáveis pretendidas, que neste caso são a polaridade do sentimento (positivo, negativo, neutro, conjunto da polaridade). A variável que foi definida como variável-alvo (*target variable*) foi a variável de conjunto (*compound),* devido a esta possuir o conjunto da polaridade o que torna possível a classificação do *tweet* como um todo;
7. De seguida, fez-se a divisão dos dados através do *data sampler,* com o objetivo de separar os dados em conjunto de treino e de teste. A divisão foi feita com 70% dos dados a serem de treino e os restantes 30% de teste, que é a medida de divisão padrão para problemas de classificação;
8. Nesta fase, fez-se a classificação do sentimento propriamente dita recorrendo a algoritmos de classificação. Estes algoritmos são: *Logistic Regression, Neural Network, Naive-Bayes, kNN e SVM;*
9. Por fim, fez-se a avaliação da classificação feita anteriormente através da funcionalidade de *Test & Score*.

No próximo capítulo vão ser demonstrados os resultados obtidos nesta fase.

Capítulo 5

5. Resultados obtidos

Neste capítulo vão ser discutidos os resultados obtidos com a análise realizada através da ferramenta *Orange 3*.

5.1 Resultados da tecnologia *Orange 3*

Nesta secção vão ser apresentados por tópicos os resultados obtidos para os dois casos de estudo definidos.

5.1.2 Resultados relativos ao caso de estudo "Mundial FIFA 2018"

5.1.2.1 Qual a equipa cujos comentários dos adeptos têm uma polaridade mais positiva?

Esta questão é respondida através da funcionalidade *Scatter Plot*. Esta funcionalidade mostra um gráfico de dispersão bidimensional da polaridade dos comentários e possui uma linha de regressão que correlaciona os dois atributos neste gráfico. Os atributos são a polaridade positiva (eixo do *x*) e polaridade negativa (eixo do *y*). O valor desta linha vai mostra o seu declive, ou seja, se possuir um valor negativo, significa que a correlação entre os atributos é negativa e variam no sentido inverso. Isto quer dizer que, se a linha possuir um valor negativo, existem mais *tweets* com polaridade positiva do que negativa.
De seguida vão ser apresentados os gráficos de dispersão dos *datasets* e também as respetivas tabelas de *test&score*, onde são avaliadas as classificações.

- Gráfico de dispersão do *dataset* da seleção francesa

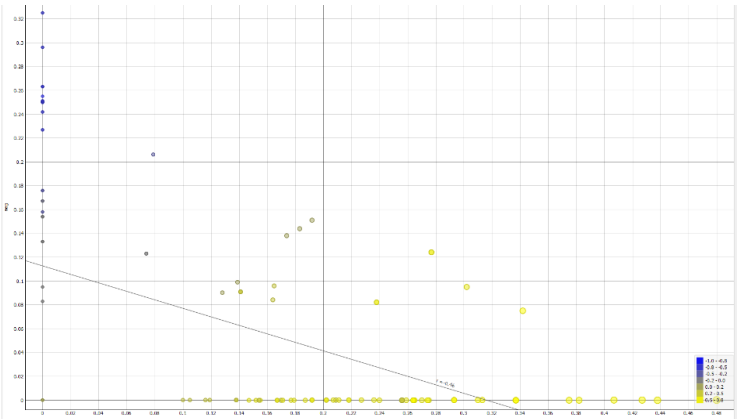


Figura 5. 1 Gráfico de dispersão seleção francesa

- *Test&Score* do *dataset* da seleção francesa

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.968	0.968	0.971	0.968
Naive Bayes	0.937	0.937	0.937	0.937
kNN	0.937	0.936	0.939	0.937
Logistic Regression	0.937	0.936	0.936	0.937
Neural Network	0.825	0.827	0.850	0.825

Figura 5. 2 Test&Score seleção francesa

- Gráfico de dispersão do *dataset* da seleção croata

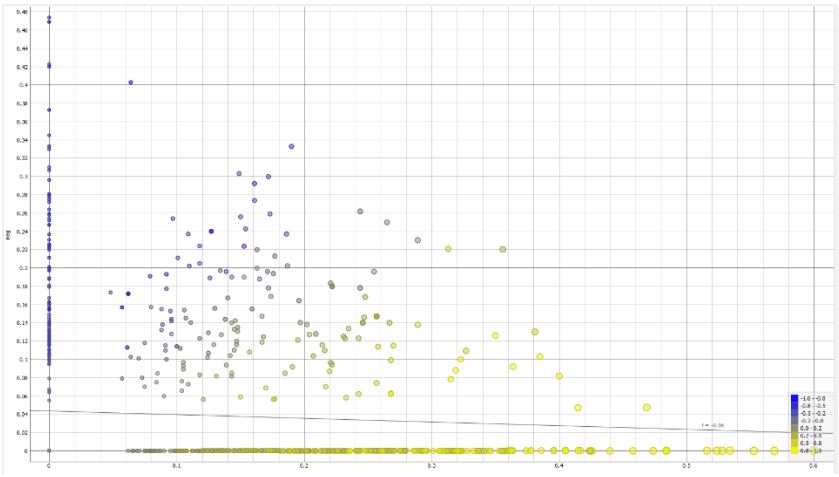


Figura 5. 3 Gráfico de dispersão seleção croata

- *Test&Score* do *dataset* da seleção croata

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.797	0.791	0.838	0.797
Logistic Regression	0.791	0.786	0.821	0.791
Neural Network	0.785	0.777	0.831	0.785
Naive Bayes	0.744	0.741	0.750	0.744
kNN	0.503	0.429	0.615	0.503

Figura 5. 4 *Test&Score* seleção croata

- Gráfico de dispersão do *dataset* da seleção belga

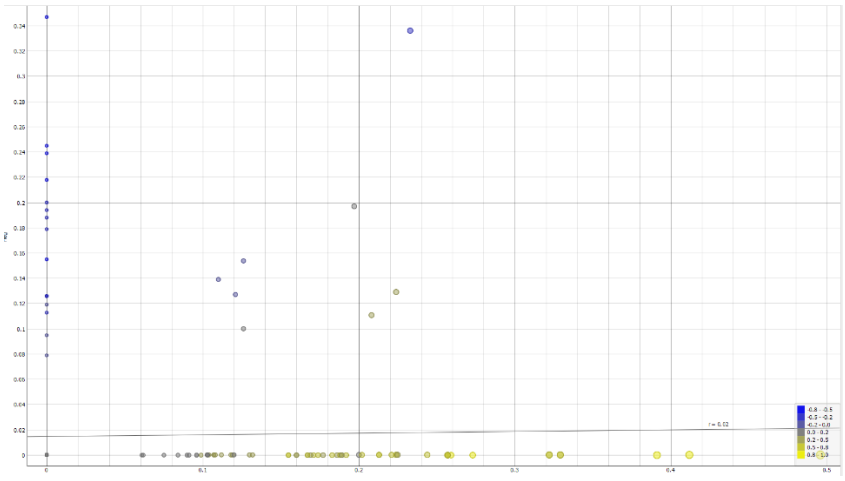


Figura 5. 5 Gráfico de dispersão seleção belga

- *Test&Score* do *dataset* da seleção belga

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.907	0.901	0.916	0.907
Naive Bayes	0.884	0.878	0.892	0.884
kNN	0.884	0.876	0.898	0.884
Logistic Regression	0.884	0.877	0.884	0.884
Neural Network	0.093	0.055	0.338	0.093

Figura 5. 6 Test&Score seleção belga

- Gráfico de dispersão do *dataset* da seleção inglesa

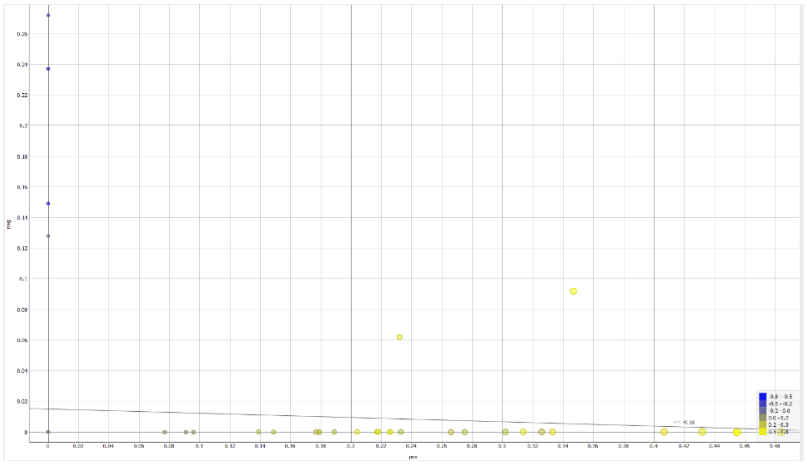


Figura 5. 7 Gráfico de dispersão seleção inglesa

- *Test&Score* do *dataset* da seleção inglesa

Evaluation Results

Method	CA	F1	Precision	Recall
Naive Bayes	0.963	0.961	0.965	0.963
SVM	0.963	0.961	0.965	0.963
Logistic Regression	0.963	0.961	0.965	0.963
kNN	0.926	0.908	0.890	0.926
Neural Network	0.852	0.812	0.789	0.852

Figura 5. 8 *Test&Score* seleção inglesa

- Gráfico de dispersão do *dataset* da seleção portuguesa

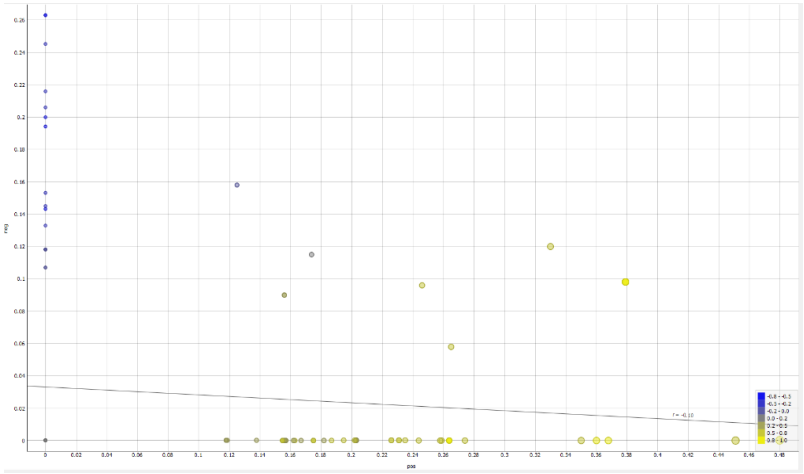


Figura 5. 9 Gráfico de dispersão seleção portuguesa

- *Test&Score* do *dataset* da seleção portuguesa

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.932	0.928	0.938	0.932
Naive Bayes	0.886	0.885	0.884	0.886
kNN	0.864	0.862	0.865	0.864
Logistic Regression	0.864	0.862	0.865	0.864
Neural Network	0.841	0.832	0.835	0.841

Figura 5. 10 *Test&Score* seleção portuguesa

Posto isto, os diferentes valores da linha de regressão para cada seleção, são os seguintes:

- Seleção da França: $r = -0.46$
- Seleção da Croácia: $r = -0.06$
- Seleção da Bélgica: $r = 0.02$

- Seleção da Inglaterra: *r = -0.08*
- Seleção de Portugal: *r = -0.10*

Com estes resultados, podemos concluir que a seleção francesa é a que tem os *tweets* com polaridade mais positiva em comparação com as restantes, ao que não é alheio o facto de ter sido a seleção vencedora da competição *FIFA World Cup 2018*. Em relação às tabelas *test&score* podemos concluir que apesar dos resultados afetados devido ao reduzido tamanho dos *datasets*, o algoritmo *SVM* é, em maior parte dos casos (quatro dos cinco *datasets*) o classificador com melhores resultados.

5.1.2.2 Como é que os jogos da competição *FIFA World Cup 2018* afetaram os comentários dos utilizadores e o tipo de emoções que estes expressaram?

A resposta para esta questão vai ser dada pela funcionalidade *Tweet Profiler* complementada com a funcionalidade *Distributions*. Vai ser dado como *input* o *dataset* específico para cada jogo entre as seleções, à exceção da seleção de Portugal, uma vez que apenas foram selecionados os jogos da fase de "oitavos-de-final", "meias-finais", final e jogo para se decidir o 3º e 4º classificado. Como *output* vão ser devolvidos os seguintes gráficos, onde se relaciona as emoções manifestadas com a correspondente frequência. São também referidos nesta secção, as tabelas de *test&score* para cada jogo respetivamente.

i) ***França VS. Croácia***

Para o caso do jogo entre a seleção da França e a seleção da Croácia, na última fase da competição, as emoções mais sentidas foram de "alegria", "surpresa" e "medo".
Estas emoções devem-se sobretudo ao facto de se tratar da última fase da competição. Podemos concluir que a emoção "alegria" vem da vitória da seleção francesa e "surpresa" pela presença da seleção croata nesta final, porque não fazia parte do grupo das seleções teoricamente favoritas a chegar a esta fase.

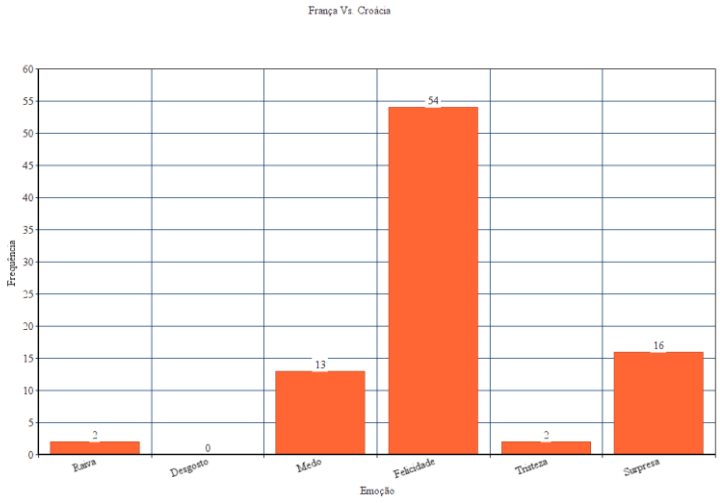


Figura 5. 11 Gráfico de barras de emoções França vs. Croácia

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.923	0.923	0.934	0.923
Naive Bayes	0.885	0.887	0.903	0.885
kNN	0.885	0.884	0.887	0.885
Logistic Regression	0.885	0.884	0.887	0.885
Neural Network	0.846	0.815	0.808	0.846

Figura 5. 12 *Test&Score* França vs Croácia

ii) ***França VS. Bélgica***

Como no primeiro caso, as emoções manifestadas nos *tweets* analisados foram "alegria" e "surpresa".

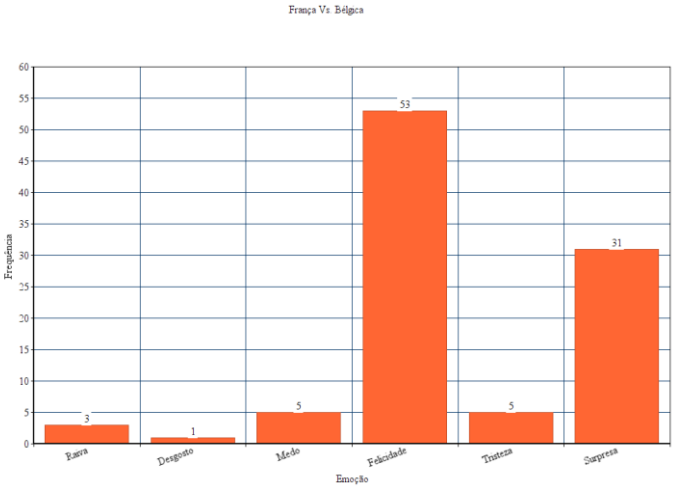


Figura 5. 13 Gráfico de barras de emoções França vs. Bélgica

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.897	0.895	0.897	0.897
kNN	0.828	0.815	0.846	0.828
Naive Bayes	0.793	0.793	0.871	0.793
Neural Network	0.793	0.749	0.711	0.793
Logistic Regression	0.759	0.749	0.776	0.759

Figura 5. 14 *Test&Score* França vs. Bélgica

iii) ***Inglaterra VS. Croácia***

No caso do jogo da seleção inglesa com a seleção croata, as emoções mais sentidas foram "medo", "alegria" e "surpresa". Neste caso, para além das emoções "alegria" e "surpresa" foi detetado um número significativo de *tweets* com o sentimento de "medo" atribuído.

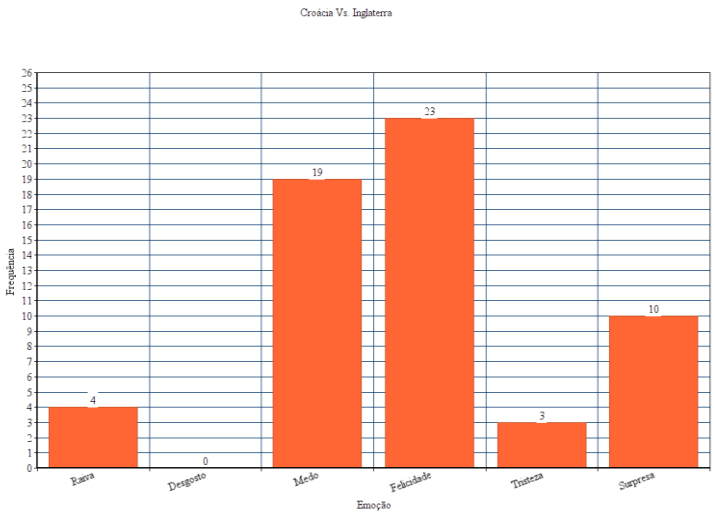


Figura 5. 15 Gráfico de barras de emoções Inglaterra vs. Croácia

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.882	0.885	0.912	0.882
Logistic Regression	0.882	0.885	0.912	0.882
Neural Network	0.824	0.826	0.882	0.824
kNN	0.824	0.827	0.840	0.824
Naive Bayes	0.765	0.770	0.809	0.765

Figura 5. 16 *Test&Score* Inglaterra vs. Croácia

iv) ***Inglaterra VS. Bélgica***

No jogo entre a seleção da Inglaterra e a seleção da Bélgica a contar para o apuramento do 3º e 4º lugar da competição, as emoções mais sentidas foram "alegria" e "surpresa".

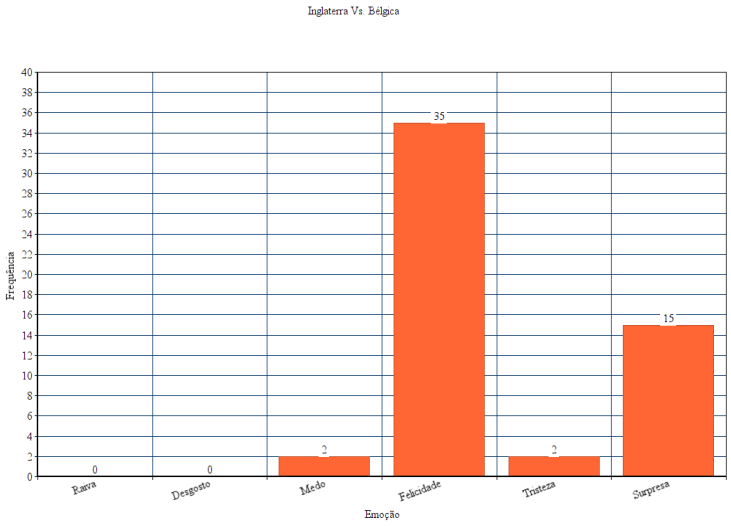


Figura 5. 17 Gráfico de barras de emoções Inglaterra vs. Bélgica

Evaluation Results

Method	CA	F1	Precision	Recall
SVM	0.938	0.937	0.944	0.938
Logistic Regression	0.938	0.937	0.944	0.938
Naive Bayes	0.875	0.873	0.900	0.875
kNN	0.812	0.806	0.864	0.812
Neural Network	0.500	0.333	0.250	0.500

Figura 5. 18 *Test&Score* Inglaterra vs. Bélgica

Por fim, construímos um gráfico comparativo das emoções sentidas em todos os jogos, como está demonstrado de seguida.

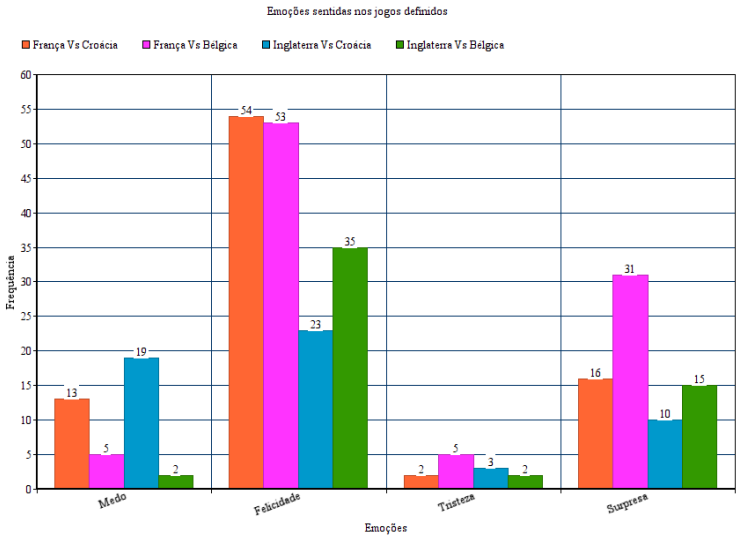


Figura 5. 19 Gráfico de barras de emoções comparativo de todos os jogos analisados

Podemos concluir que os comentários mais frequentes são os que têm emoção de "felicidade" associada. Logo a seguir estão, com uma frequência menor, os *tweets* com a emoção de "surpresa" associada. Em relação aos resultados apresentados nas tabelas de *test&score*, podemos concluir que o classificador *SVM* volta a ser aquele que apresenta melhores resultados.

5.1.3 Resultados relativos ao caso de estudo da transferência do jogador Cristiano Ronaldo do Real Madrid CF para a Juventus FC

5.1.3.1 Como é que a transferência do jogador Cristiano Ronaldo do Real Madrid para a Juventus afetou os comentários dos utilizadores afetos aos dois clubes?

Esta questão foi respondida, como foi o caso da segunda questão de pesquisa do primeiro caso de estudo, através da funcionalidade *Tweet Profiler* complementado com a funcionalidade

Distributions. Vai ser dado como *input* o *dataset* específico de cada clube relacionado com o jogador português e devolvido como *output* os seguintes gráficos onde, que como foi dito anteriormente, se relacionam as emoções sentidas com a frequência correspondente. São referidas de igual forma, as respetivas tabelas de *test&score*.

i) No caso do *dataset* que relaciona o jogador português com a equipa da Juventus a emoção predominante é a de "surpresa". Isto pode ser explicado pelo facto de que esta transferência para a equipa italiana apanhou muita gente de surpresa e isso ficou expresso nos *tweets*.

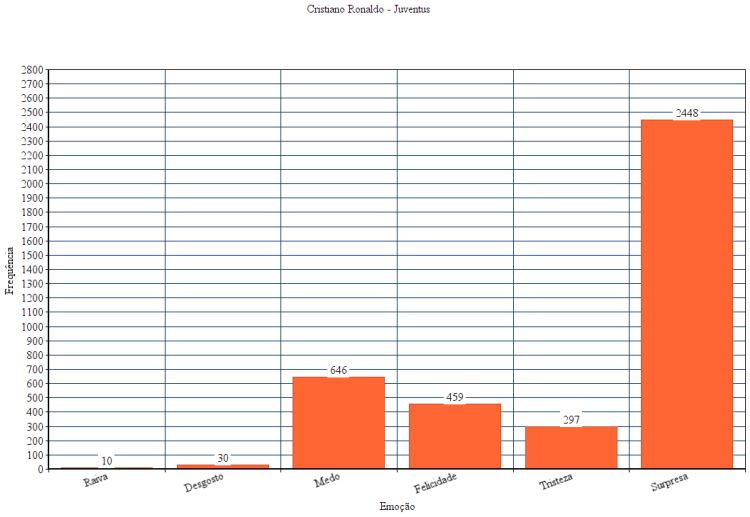


Figura 5. 20 Gráfico de barras de emoções relativos ao jogador Cristiano Ronaldo e a Juventus

Evaluation Results

Method	CA	F1	Precision	Recall
kNN	0.773	0.774	0.837	0.773
Neural Network	0.772	0.774	0.833	0.772
Logistic Regression	0.772	0.774	0.832	0.772
Naive Bayes	0.759	0.759	0.795	0.759
SVM	0.435	0.403	0.480	0.435

Figura 5. 21 *Test&Score* Cristiano Ronaldo - Juventus

ii) No caso do *dataset* relacionada com a equipa do Real Madrid, a emoção predominante também é de "surpresa". No entanto, foi identificado um número significativo de *tweets* com os sentimentos de "tristeza" e "medo" associados. Uma razão plausível para este facto é o de que a equipa espanhola ficou sem o que é considerado o melhor jogador do mundo e os adeptos afetos à equipa expressaram esses sentimentos de preocupação nos *tweets* obtidos.

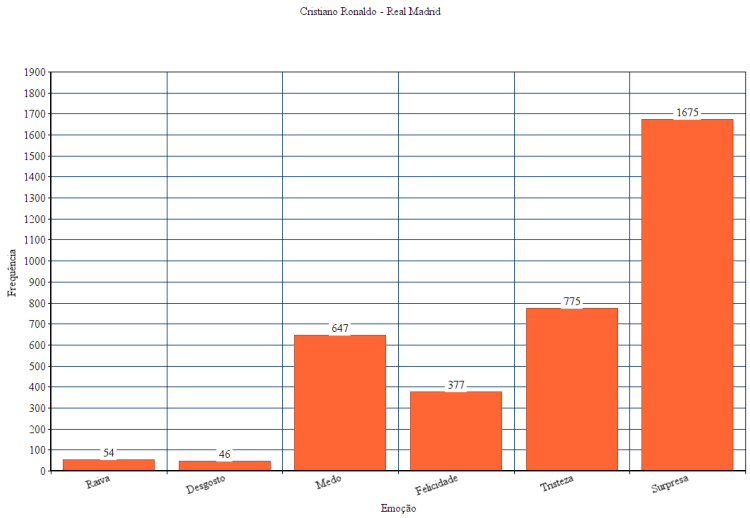


Figura 5. 22 Gráfico de barras de emoções relativos ao jogador Cristiano Ronaldo e o Real Madrid

Evaluation Results

Method	CA	F1	Precision	Recall
Naive Bayes	0.707	0.647	0.756	0.707
Logistic Regression	0.705	0.632	0.798	0.705
kNN	0.700	0.632	0.754	0.700
Neural Network	0.645	0.599	0.615	0.645
SVM	0.485	0.493	0.505	0.485

Figura 5. 23 *Test&Score* Cristiano Ronaldo – Real Madrid

De seguida mostramos um gráfico comparativo com as emoções sentidas em ambos os *datasets*.

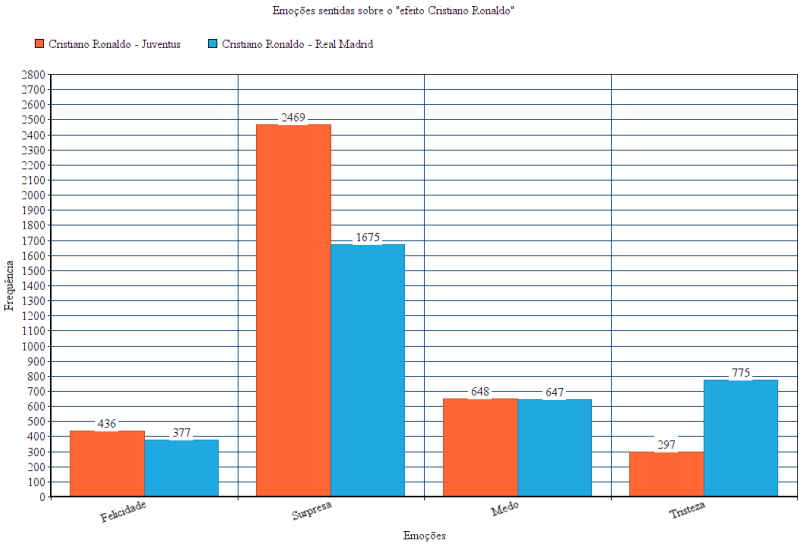


Figura 5. 24 Gráfico de barras de emoções comparativo dos *datasets* que relacionam o Cristiano Ronaldo com a Juventus e Real Madrid

Podemos concluir que a emoção mais sentida em ambos os casos foi a emoção "surpresa", apesar de existir também uma percentagem significativa de *tweets* com a emoção "medo" associada e também, por parte do *dataset* relativo ao Real Madrid CF, bastantes *tweets* com a emoção "tristeza" atribuída e isso, deve-se em grande parte, ao facto de terem perdido uma peça fulcral na sua equipa. Relativamente aos resultados das tabelas de *test&score* podemos concluir que, desta vez os valores são de tamanho mais reduzido em relação ao primeiro caso de estudo, devido ao facto de que o tamanho dos *datasets* relativos ao jogador Cristiano Ronaldo serem de maior dimensão. Posto isto, os classificadores com melhores resultados foram, para o caso da relação "Cristiano Ronaldo – Juventus", o classificador *kNN* e para o caso da relação "Cristiano Ronaldo – Real Madrid", o classificador *Naive-Bayes*.

Capítulo 6

6. Conclusão e trabalho futuro

6.1 Conclusão

As temáticas da mineração de texto e da análise de sentimentos continuam numa fase de grande crescimento e são cada vez mais procuradas nos mais diversos contextos, como por exemplo, o empresarial e académico.
Existem atualmente bastantes projetos e trabalhos de pesquisa no contexto da análise de sentimentos em redes sociais, nomeadamente no *Twitter*. No entanto, esses mesmos projetos discutem problemas e fazem análises diferentes uns dos outros. O que se pretendeu fazer com esta dissertação foi preparar um sistema simples e intuitivo, que estivesse preparado para responder a diferentes tipos de problemas relacionados com a Mineração de Texto e, consequentemente, a Análise de Sentimentos.

Contudo, foram encontradas algumas dificuldades na realização deste projeto relativamente ao primeiro caso de estudo. Uma delas foi a inflexibilidade da rede social usada, *Twitter,* que impossibilitou a extração de *tweets* que tivessem sido publicados há mais de duas semanas. Devido a este contratempo, os *datasets* adquiridos são de tamanho bastante inferior àquele que se pretendia obter. Este facto prejudicou alguns dos resultados do modelo construído, nomeadamente os resultados dos algoritmos de classificação, que com *datasets* maiores, poderiam ter sido obtidos resultados mais esclarecedores. Igualmente na deteção das emoções específicas para cada *dataset* dos jogos analisados, foram encontradas dificuldades relacionadas com o tamanho reduzido dos *datasets*. Isto veio complicar a realização dessa tarefa, já que as emoções detetadas foram bastantes similares de *dataset* para *dataset*.

Já no segundo caso de estudo, devido ao acontecimento ter ocorrido mais recentemente, não se encontraram dificuldades ao nível temporal da extração dos *tweets*, possibilitando assim uns resultados mais esclarecedores. Apesar deste facto, foram encontradas algumas dificuldades. Dificuldades essas que estão diretamente relacionadas com as acusações de alegada violação de que o jogador Cristiano Ronaldo foi alvo, tendo isso afetado ambos os *datasets,* na medida em que está presente nestes uma série de *tweets* relacionados precisamente com essas acusações.

Apesar disto, foi construído um modelo computacional pronto para dar resposta aos mais variados tipos de problemas no âmbito da análise de sentimentos.

6.2 Trabalho Futuro

Após concluída a realização deste projeto, existem ainda algumas lacunas a preencher, tanto para melhoramento do sistema construído, como para a realização de uma análise de sentimentos em outras áreas de estudo.
De seguida ficam algumas dessas ideias que achamos que podem ser de interesse realizar:

- Realização de uma análise sobre outros tópicos pertinentes, como por exemplo outras competições desportivas, eleições num determinado país ou até outro tipo de eventos de importância;
- Implementação de uma análise de sentimentos numa outra rede social com uma alta taxa de utilizadores, como por exemplo o *Facebook;*
- Melhoramento do sistema construído como o intuito de um uso mais facilitado, com a realização de atualizações.

Referências

[1] MonkeyLearn, “MonkeyLearn,” [Online]. Available: https://monkeylearn.com/sentiment-analysis/ .

[2] Carlos Augusto S. Rodrigues, et al. “Mineração de Opinião / Análise de Sentimentos”.

[3] R. Joshi, “Exsilio Solutions,” [Online]. Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

[4] David Zimbra, et al. [2018] [Online].

[5] [Online]. Available: https://www.disruptiveadvertising.com/social-media/be-in-the-know-2018-social-media-statistics-you-should-know/.

[6] Vishal Gupta, et al. “A Survey of Text Mining Techniques and Applications,” JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009, vol. 1, 2009.

[7] Jiakang Chang, et al. “Foster,” 2018. [Online]. Available: https://www.fosteropenscience.eu/content/text-mining-101.

[8] Christian Aranha, et al. “A Tecnologia de Mineração de Textos,” 2006.

[9] Sarah E. Shukri, et al. “Twitter sentiment analysis: A case study in the automotive industry,” 2015.

[10] Elif Uysal, et al. “Sentiment Analysis for the Social Media: A Case Study for Turkish General Elections,” 2015.

[11] Wu He, et al. “Social media competitive analysis and text mining: A case study in the pizza industry,” 2013.

[12] “Extreme Tech,” [Online]. Available: https://www.extremetech.com/computing/252987-python-tops-list-2017s-popular-programming-languages.

[13] [Online]. Available: https://www.predictiveanalyticstoday.com/compare/orange-data-mining-vs-rapidminer-starter-edition-vs-knime-analytics-platform-community/ .

[14] C.J. Hutto, et al. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”.

[15] S. SHARMA. [Online]. Available: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6.

[16] Avinash Sharma, 2017. [Online]. Available: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0.

[17] [Online]. Available: https://pythonmachinelearning.pro/text-classification-tutorial-with-naive-bayes/.

[18] [Online]. Available: https://www.wattpad.com/115821129-algoritmos-de-aprendizagem-de-m%C3%A1quina-knn-nearest.

[19] M. de Vries, “Machine Learning for Sentiment Analysis,” 2017.

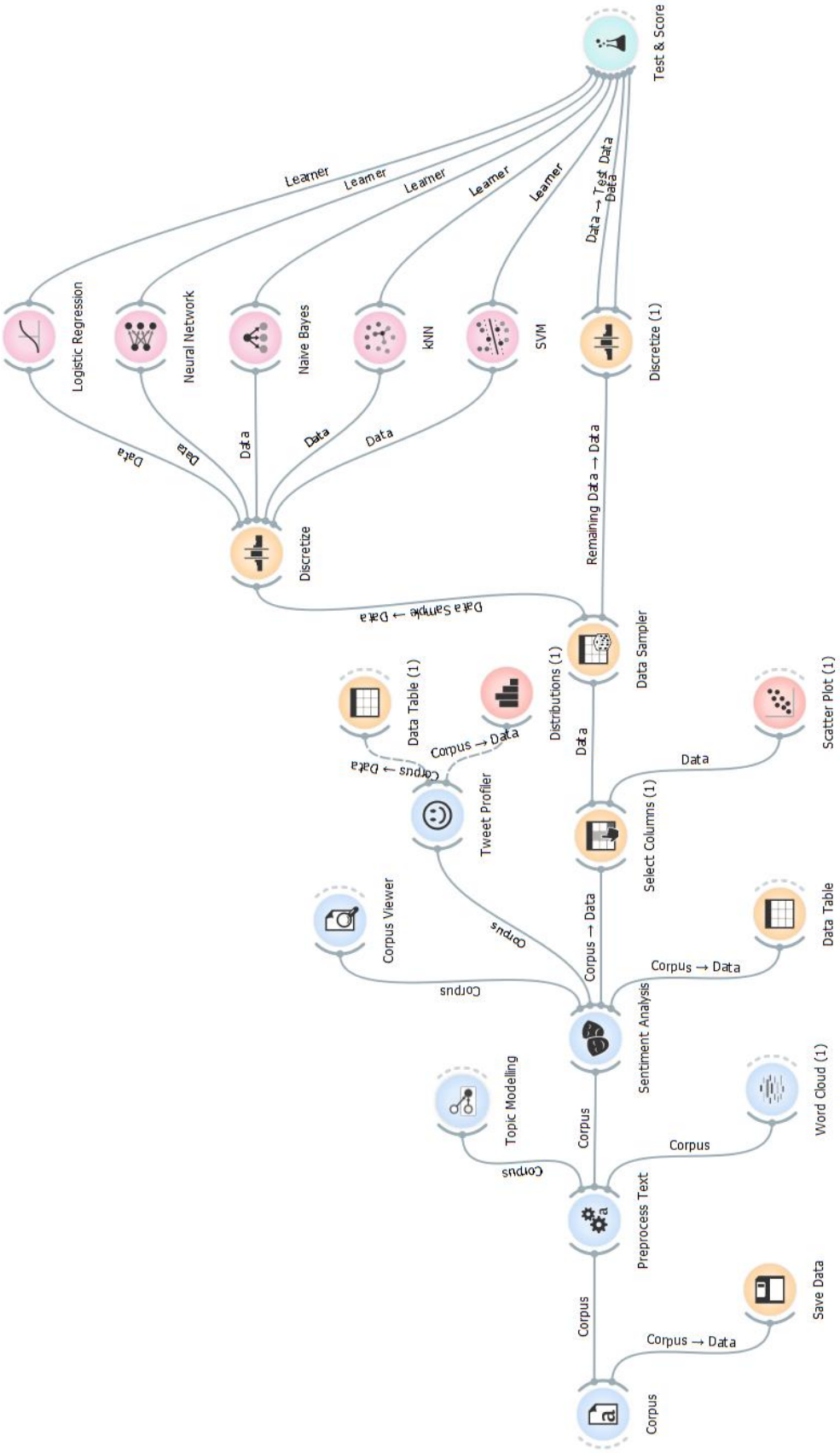
[20] Nurulhuda Zainuddin, et al. “Sentiment analysis using Support Vector Machine,” 2014.

Anexos

Anexo A – Código *Python* para o *download* dos *datasets*

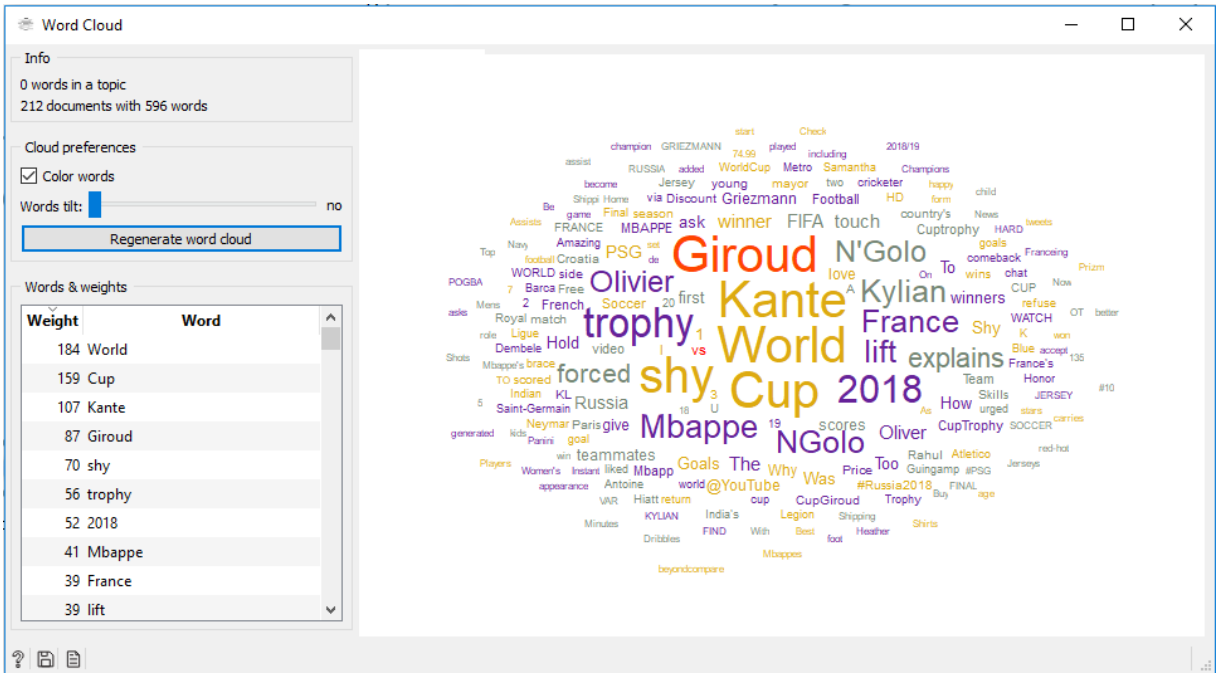
```
import tweepy
import csv

import unicodedata


####inserir credenciais
from nltk import data
from textblob import TextBlob

consumer_key = '######################'
consumer_secret = '##########################'
access_token = '#################################'
access_token_secret = '###############################'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)


# criar ficheiro .csv
csvFile = open('FileName.csv', 'a')
#Use csv Writer
csvWriter = csv.writer(csvFile, delimiter=',')



#inserir query para pesquisa
for tweet in tweepy.Cursor(api.search, q="Insert Query Here", count=10000,
lang="en").items():
    if (not tweet.retweeted) and ('RT @' not in tweet.text):
        print (tweet.created_at, tweet.text)

    #for row in row:
        csvWriter.writerow([tweet.created_at, tweet.text.encode('ascii',
'ignore')])

    # print("")
```

Anexo B – Workflow do Orange 3

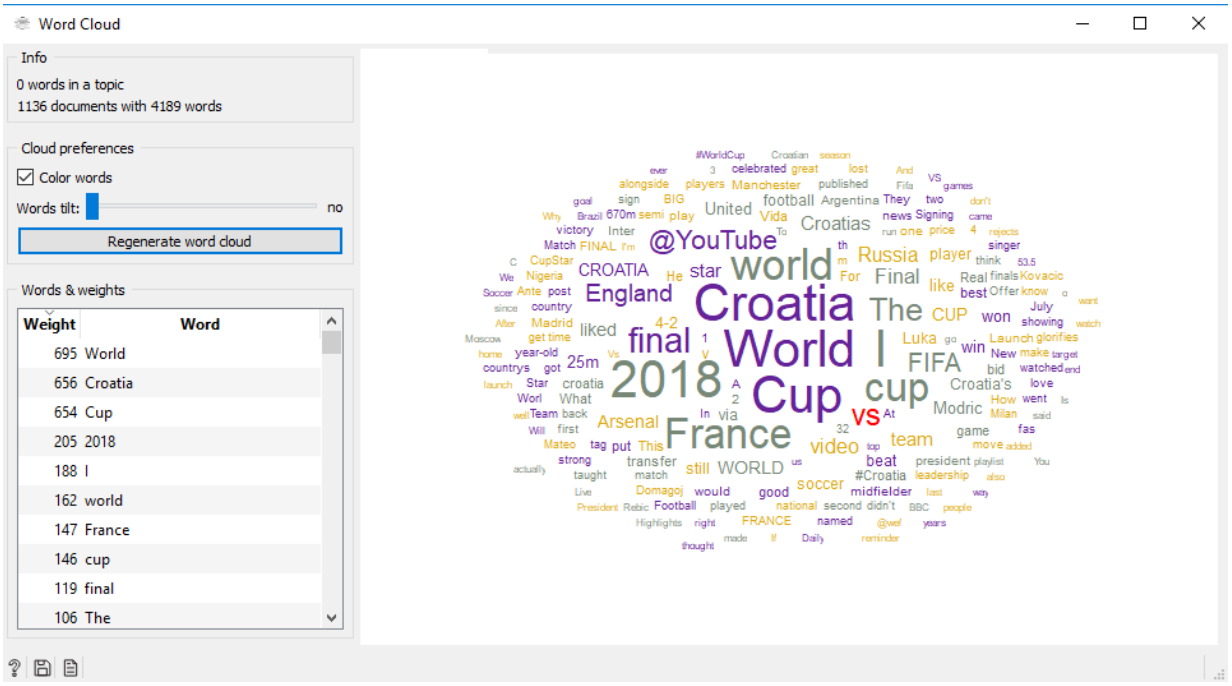


Anexo C – Word Clouds

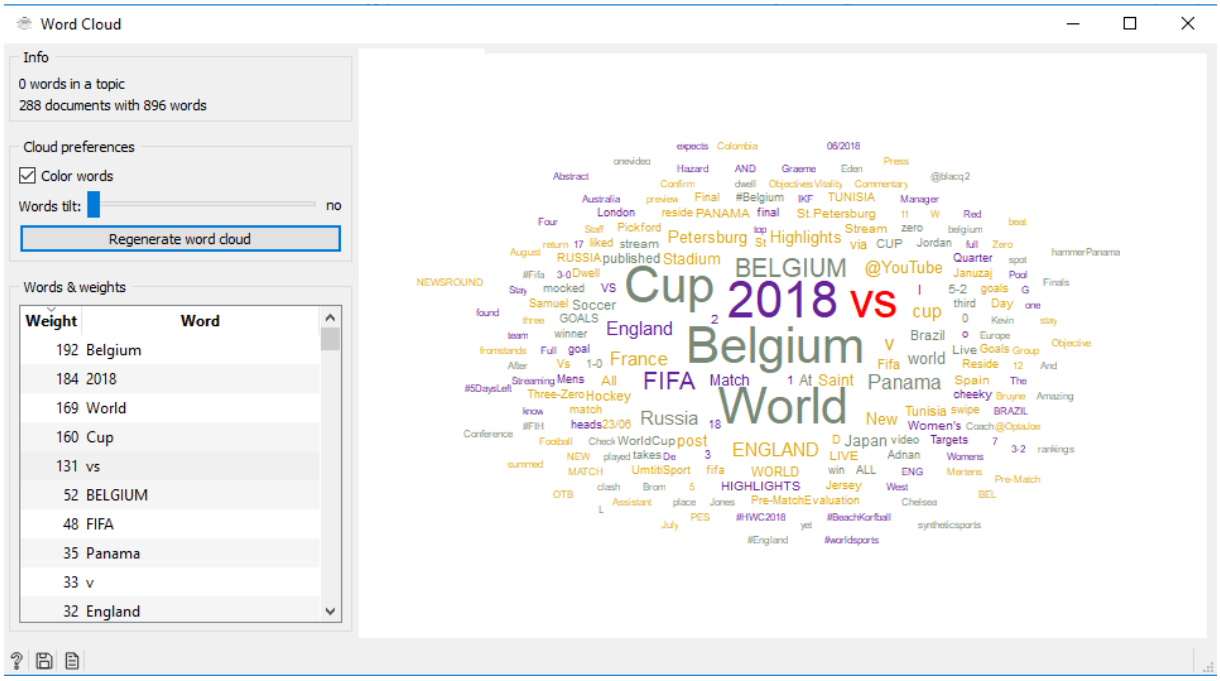
i) *Word Cloud dataset* Seleção da França



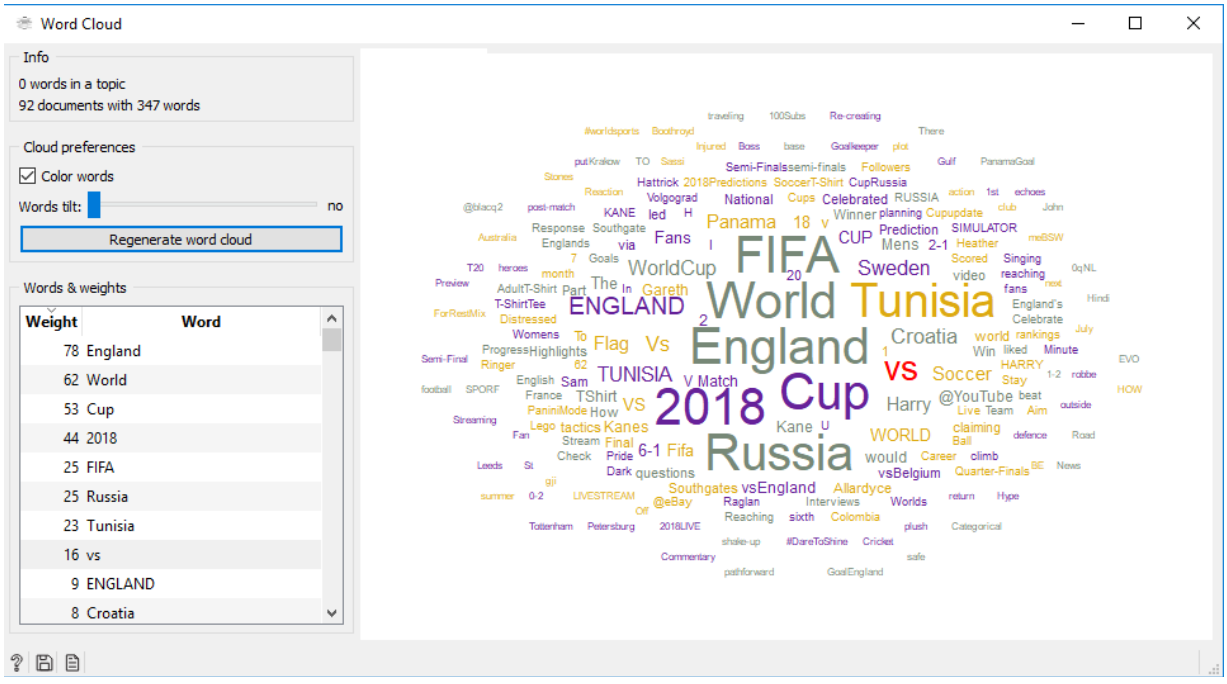
ii) *Word Cloud dataset* Seleção da Croácia



iii) *Word Cloud dataset* Seleção da Croácia



iv) *Word Cloud dataset* Seleção da Inglaterra



v) *Word Cloud dataset* Seleção da Portugal